OPEN ACCESS

Chat-IRB? How application-specific language models can enhance research ethics review

Sebastian Porsdam Mann ,[1,2,3] Jiehao Joel Seah ,[3] Stephen Latham,[4] Julian Savulescu,[3,5] Mateo Aboy,[6] Brian D Earp [3,5]

[1]Centre for Advanced Studies in Bioscience Innovation Law (CeBIL), Faculty of Law, University of Copenhagen, Copenhagen, Denmark
[2]Faculty of Law, University of Oxford, Oxford, UK
[3]Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[4]Yale Interdisciplinary Center for Bioethics, Yale University, New Haven, Connecticut, USA
[5]Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, Oxford, UK
[6]Centre for Law, Medicine, and Life Sciences (LML) & Centre for Intellectual Property and InformationLaw (CIPIL), Faculty of Law, University of Cambridge, Cambridge, UK

Correspondence to
Professor Julian Savulescu; julian.savulescu@uehiro.ox.ac.uk

Received 13 February 2025
Accepted 3 July 2025

To cite: Porsdam Mann S, Seah JJ, Latham S, *et al*. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jme-2025-110845

ABSTRACT

Institutional review boards (IRBs) play a crucial role in ensuring the ethical conduct of human subjects research, but face challenges including inconsistency, delays, and inefficiencies. We propose the development and implementation of application-specific large language models (LLMs) to facilitate IRB review processes. These IRB-specific LLMs would be fine-tuned on IRB-specific literature and institutional datasets, and equipped with retrieval capabilities to access up-to-date, context-relevant information. We outline potential applications, including pre-review screening, preliminary analysis, consistency checking, and decision support. While addressing concerns about accuracy, context sensitivity, and human oversight, we acknowledge remaining challenges such as over-reliance on artificial intelligence and the need for transparency. By enhancing the efficiency and quality of ethical review while maintaining human judgement in critical decisions, IRB-specific LLMs offer a promising tool to improve research oversight. We call for pilot studies to evaluate the feasibility and impact of this approach.

INTRODUCTION

Research involving human participants has the potential to yield invaluable knowledge that can advance medical practice, improve health outcomes, and inform us about societal practices and human behaviour. However, history has shown that such research also carries risks for participants who may be subjected to exploitation or unjustifiable harm if adequate protections are not in place.[1 2] Consequently, a robust framework of ethical norms, codes of conduct, regulations, and laws has emerged to guide and govern human subjects research.[3–7]

A key component of the ethical oversight infrastructure (ie, Human Research Protection Programs) is the institutional review board (IRB), also known as a research ethics committee. Before research involving human participants can proceed, the study protocol must undergo review and approval (and potentially also monitoring post-approval) by an IRB to ensure regulatory compliance, sound scientific design (ie, to prevent harm to participants arising from scientific design), and ethical acceptability. In theory, the IRB review process is an essential safeguard to protect research participants from excessive risk and unjustifiable harm. In practice, however, it suffers from several well-documented problems and limitations.[8] The net result of these problems is an ethics review process often marked by delays, unpredictability, and insufficient accountability—issues which can impede valuable research without necessarily improving protections for human participants.

For example, studies have found wide variability in how different IRBs interpret regulations, make value judgements, and reason through ethical issues.[9 10] There are also significant discrepancies in the levels of scrutiny applied, with some protocols (based on the IRB's determination of the study's level of 'minimal risk' and its risk appetite) undergoing 'full board' review while very similar studies are classified as 'exempt' or eligible for 'expedited' review. This inconsistent application of standards is compounded by an apparent lack of transparency in many cases (ie, IRB deliberations and decisions appear to be shrouded in secrecy due to the lack of adequate rationale in its communications to researchers, or from the absence of sufficient and accurate documentation of the IRB's reviews).[11] And there is often no robust mechanism for an institutional memory of past IRB discussions and decisions.

In light of these limitations, it has recently been proposed to make use of advances in generative artificial intelligence (AI), in particular large language models (LLMs), as a *first pass* or *adjunct* screening tool to facilitate the speed with which IRBs can assess research proposals.[12 13] While general LLMs—such as those employed in OpenAI's GPT, Meta's Llama, or Anthropic's Claude series—have shown initial promise and potential in this area,[12 13] their use as screening tools is also attended by risks and potential ethical concerns. Ethical review is a paradigmatic case of an area in which *human* oversight and judgement are widely believed to be of paramount importance.[14] This belief is likely strengthened by an awareness of known issues with LLMs related to accuracy, bias, explainability, consistency, reproducibility, and the potential for LLMs to 'hallucinate' (ie, the ability of LLMs to generate plausible-sounding outputs that are not directly grounded in factual or training data due to the generative AI models' capacity to create, synthesise, or extrapolate information probabilistically based on its learnt patterns).[15 16] Moreover, most IRBs, unlike most off-the-shelf LLMs, are institution specific: there may be particular institutional requirements that need to be followed based on local factors, including local regulations, or specific concerns of the local population, which are germane to ethical review decisions.

For these reasons, any proposal to cede or delegate important aspects of ethics oversight to AI is understandably controversial. We readily acknowledge these concerns. One way to address them, as has previously been proposed, is to use these systems not as a replacement but rather as a supplement to

human review.[12] In this paper, we build on this suggestion by proposing the augmentation of human review with application-specific IRB LLMs designed and validated specifically for this purpose. This design can be achieved through tried-and-tested technical means such as application-specific fine-tuning, retrieval-augmented generation (RAG), and prompt engineering.[17] To the extent that the design, implementation, and rigorous validation of an application-specific IRB AI is successful, the resulting LLM will have access to additional and higher-weighted knowledge specific to individual IRBs, which partially addresses concerns related both to accuracy and to the need for context sensitivity. Such purpose-built LLMs could handle pre-submission checks and initial reviews, freeing up time and attention for more complex, sensitive, or high-risk cases; however, difficult judgement calls and final determinations would still rest with human decision-makers.

In the sections below, we develop this argument in greater detail. We begin by providing background information on the need for, as well as known problems with, IRB review, as well as briefly detailing existing proposals for the use of LLMs to address these problems. Next, we provide a concise technical overview of LLMs generally and of application-specific IRB LLMs specifically, illustrating their potential with examples from the burgeoning literature of personalised LLMs for bioethical applications. Building on these technical foundations, we go on to present an operational vision for IRB-specific LLMs and walk through how they could be integrated into existing workflows to improve speed, consistency, and quality. Finally, we consider some of the key limitations and open questions related to our proposal, charting an agenda for future research and implementation efforts.

PROBLEMS WITH ETHICAL REVIEW OF HUMAN SUBJECTS RESEARCH

The primary function of IRBs is to ensure that research protocols comply with applicable laws and regulations, have sound scientific design, and are ethically acceptable. This involves a careful assessment of the risks and anticipated benefits of the proposed research, as well as a determination of whether the rights and welfare of participants are adequately protected. IRBs also have a responsibility to ensure that consent will be obtained in a manner that is free from coercion or undue influence, and that participants are presented with adequate information about the study.

While the requirement of IRB review has certainly played a significant role in the furtherance of research ethics, the process can subject proposed research to significant, although unnecessary, delays which may in turn slow down scientific progress and hinder advances in medical treatment.[18]

These delays stem from multiple sources. Many IRB offices face significant challenges in institutional funding and staffing—which can be traced to (historical) structural decisions that fail to allocate dedicated or sufficient financial support—resulting in inadequate human resources and overworked offices with extended turnaround times.[19 20] In our own experience, poor quality submissions[i] exacerbate delays, often lacking key details, containing discrepancies, missing documents, or using excessive technical jargon in participant-facing materials. Delayed researcher responses to IRB queries further prolong the process. Unclear and complex IRB communications, especially regarding required modifications and their rationales, can lead to misunderstandings. Research in jurisdictions with complicated legal landscapes—such as the USA, with its federal and state laws—or in international multi-site studies adds complexity. Critically, the lack of institutional memory—specifically, the absence of a searchable database of past IRB decisions and outcomes—hinders efficient reviews and contributes to variability in IRB interpretations and decisions.[21 22] Additionally, inadequate procedures for the secure handling of participants' data often complicate the approval process, further extending the time required for IRB review and approval.

In light of these challenges, there is a clear need for innovative solutions that can enhance the speed, consistency, and quality of the IRB review process, as well as address IRBs' staffing and resource constraints resulting from institutional funding decisions. One promising approach is to leverage the power of generative AI, particularly LLMs, to assist with the initial screening and evaluation of research protocols.

GENERATIVE AI FOR IRB REVIEW

LLMs are general-purpose AI models whose primary function is to predict the next token (a short word or a fragment of a word) in a sequence of natural language. Recent advances in data availability, hardware, and model architecture have greatly expanded the capacities of multimodal generative AIs to process text and other modalities, such as sound, image, and video.[23] At the time of writing, state-of-the-art LLMs are trained on huge corpora of text and show significant performance across a wide variety of information tasks.[24] Recently, the introduction of reasoning models or 'reasoners' (eg, OpenAI's o-series, Anthropic's Claude Sonnet and Opus with their 'extended thinking' mode, DeepSeek R1) now confers LLMs the capability to spend more time processing a request, breaking it down into a series of steps before providing a response.[ii] This 'Chain-of-Thought' (CoT) technique—an automated process which directs the AI to approach the problem/user's prompt in a stepwise fashion—can be used to generate outputs that purportedly surpass traditional LLMs (eg, GPT-4o) significantly, in tasks that are reasoning heavy. In essence, when the LLM is directed to 'think'—and when it 'thinks' longer—the better its responses would generally be. The development of reasoners has also led to advancements in agentic AIs—AIs which, on being given a goal, have the ability to achieve that goal autonomously—that is, break it down into necessary steps and executing these – with no further human input.[iii]

LLMs are increasingly used in biomedical contexts, including for interpreting and generating clinical notes and patient reports,

[i] 1. By "poor quality submissions", we refer specifically to the quality of the IRB application form itself – eg, how clearly and completely the required information is presented – which often includes details of the research protocol/methodology. "[Q]uality" here does not refer to the scientific quality or methodological rigour of the protocol, the review of which primarily falls under the purview of scientific review committees rather than IRBs – except where poor scientific design may expose potential participants to unnecessary risks and/or harms.

[ii] 2. See generally: https://openai.com/index/learning-to-reason-with-llms/

[iii] 3. Novel real-world applications of AI agents include OpenAI's Operator and Deep Research. Operator is an agent that browses the internet to autonomously execute human tasks (eg, filling up forms, ordering groceries), while Deep Research is an agent that performs 'multi-step' research by aggregating and 'synthesising' content from various online sources. Leveraging OpenAI's o3 reasoning model, it interprets, "reasons", and analyses the gathered information to generate, within minutes, a comprehensive fully cited report with reportedly high-quality references (eg, an up-to-date detailed report on gene therapies that have secured regulatory approval for condition X).

to assist in limited areas of diagnostics and treatment, as well as in various administrative and research-oriented tasks.[25] Consequently, the idea of using LLMs to screen protocols submitted to IRBs has been raised.[12] Indeed, an early study demonstrated the potential of this idea across a series of case studies using four different LLMs, each of which was able to pick up on several ethical issues associated with each of seven case studies.[12] However, the study also noted several key omissions, concluding that human oversight continues to be important and that LLMs should be used only as adjunct screening tools.[12]

While the use of LLMs as screening tools prior to a human review is promising, it is also associated with several practical and ethical concerns. Off-the-shelf LLMs trained on generic datasets sometimes lack the domain-specific knowledge and contextual understanding necessary to make nuanced judgements about the ethical acceptability of research protocols. They may also be prone to biases and blind spots that could lead to unfair or unreliable assessments, make mistakes, and miss important areas of potential ethical concern.[12] Moreover, the 'black box' nature of traditional LLMs' internal mechanisms makes it difficult to interrogate their reasoning *a priori* and ensure that their potential outputs will align with relevant ethical principles and regulatory requirements. However, with the advent of reasoning models, we recognise that newer LLMs can now provide a degree of interpretability to their responses (which, in the process, also ameliorates concerns regarding hallucinations). By making its chain of thought available to users, the LLM's reasoning processes can be interrogated through human evaluation and validation of the logic, veracity, depth, and breadth of reasonings.

In addition to these practical issues, many people may experience general discomfort with the notion of ethical review carried out—perhaps to any extent—by AI systems. Ethical deliberation is a complex process involving not only pure analysis and computation but also qualities such as situational awareness, empathy, and nuanced balancing of competing ethical concerns. There is a legitimate worry that relying too heavily on AI systems, even as a screening or support tool, could lead to a diminishment of these essential human qualities in the review process. IRB members may feel less inclined to engage deeply with the ethical dimensions of research proposals if they believe that an AI system has already done the heavy lifting for them. Over time, this could lead to a gradual erosion of the ethical skills and sensitivities that are so crucial to the integrity of the review process.

Unlike other areas where AI is being deployed, such as medical diagnosis or financial risk assessment, the judgements made by IRBs are not purely empirical or technical in nature. They involve weighing competing moral considerations, interpreting abstract principles in light of concrete circumstances, and making difficult trade-offs between individual rights and interests, and potential societal benefits. These are quintessentially human tasks that require a deep understanding of the ethical frameworks and values that undergird the research enterprise. Delegating such tasks to AI systems, even partially, may be seen as an abdication of moral responsibility on the part of IRBs.[26]

Another critical concern is the potential inability of LLMs to fully account for the local institutional and cultural contexts in which IRBs operate. While the basic principles of research ethics—such as respect for persons, beneficence, and justice—are widely applicable, their interpretation in practice can vary significantly depending on the specific setting and population involved. For example, what constitutes an acceptable level of risk or an appropriate consent process may differ depending on the cultural norms, educational backgrounds, and socioeconomic conditions of research participants. Similarly, the institutional mission, resources, risk tolerance, and policies and procedures of a particular IRB may shape its approach to protocol review in ways that are not easily captured by a generic AI model.

This context-specificity is one of the reasons why IRBs are typically localised within particular institutions or research communities, rather than being centralised at a national or international level. Especially in the case of IRBs overseeing social[1] and behavioural research, they are expected to have a deep understanding of the local research landscape and to be attuned to the needs and concerns of the populations they serve.

Thus, the current situation may be summarised as follows. IRB review serves an important function, yet the conditions necessary for this function to be effectively fulfilled are often not available to those charged with carrying out IRB review. As a result, the quality of reviews may often be lacking, while much beneficial research is delayed and otherwise hindered. The use of LLMs to assist in pre-review screening (eg, for the completeness of the IRB application, informational consistency across the document, and alignment with ethical norms and legal requirements prior to ethical review) has been proposed as a means of addressing this delay. While promising, such use is attended by significant concerns partially related to accuracy and partially related to more general worries about the nature of ethical reasoning and the need for review that is sensitive to the local context.

In the sections that follow, we suggest that the use of application-specific LLMs trained on or given access to previous local IRB queries (ie, questions raised by the IRB to researchers), as well as IRB committee discussions and decisions, institutional and IRB-specific policies and procedures, and local regulations, could both increase accuracy and significantly ameliorate the bulk of these concerns.

APPLICATION-SPECIFIC IRB LLMS

LLMs achieve their general-purpose capabilities due to the vast and diverse datasets used in their initial training. However, when high performance is desired within a specific domain rather than for general tasks, the effectiveness of LLMs can be significantly enhanced through a process known as fine-tuning. Fine-tuning involves further training an already pre-trained LLM on a targeted, specialised dataset relevant to the specific area of interest. This process refines the model's understanding by adjusting a subset of its internal weights to align more closely with the patterns, terminology, reasoning, knowledge base, and nuances of the specialised data. As a result, fine-tuned AI models become more adept at generating accurate and contextually relevant outputs within their application-specific domains, typically surpassing the performance of general-purpose models in tasks such as medical diagnosis, legal analysis, and other field-specific applications.[27 28]

Application-specific fine-tuning helps reduce hallucinations in LLMs by aligning the model's outputs more closely with domain-specific knowledge. This can be further enhanced through RAG: a strategy that grounds the model's outputs in external, reliable data sources (ie, domain-specific reference information).[29] RAG works by using embeddings, which are numerical representations that encode text into structured vectorised forms, allowing the model to perform efficient computational searches across vast datasets. This structured approach enables the model to locate and retrieve specific, contextually relevant information during text generation, acting as an external memory or knowledge base. By grounding the model's responses in actual data, RAG helps ensure that outputs are more accurate and reliable. Practical examples of RAG include company AIs that access

internal documents to provide precise answers to user queries, thereby reducing the risk of generating incorrect or fabricated information.[30]

We use the term 'application-specific LLMs' to refer to LLMs, which have been (1) pre-trained, (2) fine-tuned for a specific application domain (ie, IRB reviews), (3) enhanced with domain-specific information using RAG, or (4) customised for a specific application through prompt engineering. The use of application-specific LLMs has the potential to address several of the specific concerns raised above with respect to the use of general LLMs in IRB review. Recently, a growing body of literature has investigated the use of a specific type of LLMs within bioethics, including LLMs that use for their training data or knowledge base information produced by, or characterising, a specific individual.[31] Examples of such *personalised* LLMs include models fine-tuned on an individual's previously published works, enabling them to produce text more in line with the style, format, and argumentation used by that individual than a base model[32 33] and similar models fine-tuned on relevant data for individual patients which might be used to predict that patient's preferences in cases of medical incapacity.[34]

Building on these preliminary examples, we envision the development of IRB-specific LLMs that are designed and validated to the unique needs and institutional contexts of individual review boards. These models would be fine-tuned on IRB-specific datasets. They would also be equipped with retrieval capabilities that allow them to access and incorporate the latest relevant information from knowledge bases established for that purpose. Note that while we focus here on individual IRBs, our proposal of using application-specific rather than generic LLMs for IRB review is also germane to centralised or supra-IRBs.[35]

The process of developing an IRB-specific LLM would involve several key steps:

1. *Data collection and curation*: the first step would be to gather a comprehensive dataset of materials relevant to the IRB's work, including IRB literature (eg, fundamentals about clinical trial design and good practices, IRB and bioethics peer-reviewed literature), past protocol submissions and related documents (eg, consent forms, study advertisements), decision letters and IRB queries (ie, IRB submission or review-related communications), meeting minutes, institutional and IRB policies and procedures, and relevant laws and regulations. This dataset would need to be carefully curated to ensure its quality, with attention to issues such as bias and confidentiality. Besides key IRB documentation, recommendations and guidelines from national ethics bodies and/or expert advisory committees (eg, national bioethics committees, ethics think tanks) could also be included in the model's database for RAG.
2. *Model selection*: the next step involves selecting an appropriate existing general LLM to serve as the base model for fine-tuning and deployment. Available options include proprietary models such as Anthropic's Claude and OpenAI's GPT and o-series, as well as open-source alternatives such as Meta's LLaMA, or DeepSeek's V3 and R1 models. The choice of model should be guided by several key considerations, including performance, scalability, and compatibility with the institution's technical infrastructure and data security requirements. Smaller, open-source models have the advantage of being deployable on local servers, offering enhanced control over data handling and partially mitigating confidentiality concerns. This local deployment option can be particularly beneficial when sensitive IRB-related data is involved, as it allows institutions to maintain tighter control over data flow and access. The final model selection should align with the IRB's institutional policies, including those outlined in its AI governance framework or policy, to ensure compliance with requirements for data processing, security, and ethical standards. Additionally, the selected model must support the IRB's specific needs, such as the ability for enterprise-quality data protection, application-specific fine-tuning, and the ability to integrate with RAG to access and use domain-specific knowledge efficiently. Ultimately, the chosen model should not only meet technical criteria but also adhere to the IRB's broader institutional and ethical mandates, ensuring responsible use of AI within the context of research ethics review.
3. *IRB application-specific fine-tuning*: the fine-tuning process can be divided into two key stages: (1) creating a general application-specific IRB LLM; and (2) further training this IRB LLM to align with the specific institutional context. In the first stage, the base model is fine-tuned on a comprehensive IRB-specific dataset, incorporating general legal standards, regulations, best practices, and relevant peer-reviewed literature. This step equips the model with a deeper understanding of the broader IRB landscape, enabling it to recognise, 'reason', and generate outputs that reflect the unique language, terminology, and ethical frameworks commonly used in IRB decision-making. In the second stage, the model undergoes additional fine-tuning to adapt specifically to the institution's context. This involves training the IRB LLM on high-quality text describing the institution's specific IRB processes, policies, historical decisions, and other relevant documentation. The goal is to imbue the model with the distinct language patterns and contextual nuances that capture the institution's decision-making criteria, ethical considerations, and procedural standards. This institution-specific training step ensures that the model's CoT/reasoning and outputs are not only aligned with general IRB practices but also tailored to reflect the unique values and operational specifics of the institution. Together, these steps enable the fine-tuned IRB LLM to produce highly relevant, context-aware outputs that closely align with both general and institution-specific decision-making frameworks, enhancing its effectiveness as a support tool in the ethical review process.
4. *RAG*: to enhance the model's ability to provide up-to-date and contextually relevant information, it would be equipped with retrieval capabilities using embeddings. This would involve creating structured vector representations of key documents in the IRB's factual knowledge base, such as past decisions, policies, and guidelines. During the generation process, the model could then efficiently search and retrieve relevant information to inform its analyses and recommendations.
5. *Testing and validation*: before being deployed in a live IRB setting, the application-specific LLM would need to undergo rigorous testing and validation to ensure its accuracy, reliability, fairness, and adherence to ethical and regulatory principles. This would involve both quantitative evaluations of the model's performance on benchmark tasks, as well as qualitative assessments by IRB staff, members, and other stakeholders–such as legal counsel–to evaluate the model's practical utility, fairness, and alignment with ethical principles. These evaluations would focus on the model's ability to: provide logical and high-quality reasoning with sufficient depth and breadth; generate accurate, contextually relevant outputs; identify potential biases; and ensure compliance with applicable laws and regulations. Feedback from these assessments should guide iterative refinements, ensuring that the model

meets the rigorous standards required for responsible use in IRB decision-making and review processes. By comparing the model's reasonings and outputs to previous human IRB decisions on similar cases, evaluators can assess the accuracy, consistency, and alignment of the AI's recommendations with established IRB standards and practices.

6. *Post-deployment audits and continuous improvement through reinforcement learning from human feedback*: after deployment, the LLM would be subject to periodic audits to monitor its performance and ensure ongoing alignment with IRB standards. This process would involve systematically comparing the LLM's reasonings and recommendations with the IRB's deliberations and final decisions. Differences between the initial LLM output and the final IRB decision, which integrates human judgement in addition to the IRB-LLM, can serve as valuable data for continuous improvement. This before-and-after analysis allows for high-quality reinforcement learning through expert human feedback, enabling the model to learn from expert corrections and adjustments made during IRB reviews. Such feedback loops are essential for refining the application-specific LLM, addressing deficiencies or biases in the model's training data, and enhancing its decision-making capabilities over time. The audit process would also incorporate direct user feedback, focusing on scenarios where human reviewers diverged from the LLM's recommendations, the frequency of these instances, and the reasons behind them. This structured feedback mechanism not only identifies the model's limitations but also guides iterative retraining, ensuring the LLM evolves to better support IRB decision-making while maintaining ethical integrity and regulatory compliance.

Once developed and rigorously validated, the IRB-specific LLM could be integrated into the review workflow in various ways to enhance the efficiency and quality of the process. For example:

1. *Pre-review screening*: the LLM (or AI agent systems, such as OpenAI's 'ChatGPT Agent' or Anthropic's 'Claude Code') could be used to autonomously screen incoming protocol submissions for completeness and potential ethical concerns, as well as propose recommended classifications (ie, 'exempt', 'expedited', or 'full board' categories). It could flag submissions that require further information (eg, missing key information in the consent form) or that raise significant concerns (eg, potential non-compliance with laws, regulations, or institutional policy; mention of vulnerable populations, and so on), allowing IRB staff to prioritise their review efforts. Similarly, the same function could also be extended to investigators to screen their draft applications, prior to IRB submission, to improve the quality and completeness of their submissions, while also flagging unnecessary jargon or potentially confusing passages of text.
2. *Preliminary review*: for protocols that pass the initial screening, the LLM/AI agent could autonomously generate a preliminary review report that: summarises the study aims and methodology, key study design and ethical considerations; identifies relevant precedents and guidelines; and provides a tentative risk–benefit assessment. This report could serve as a *starting point* for the actual human review by the IRB member(s) or discussions at convened IRB meetings.
3. *Consistency checking*: the LLM, powered by a reasoning model, could be used to evaluate the IRB's proposed clarifications, modifications, or decisions with its past precedents and with relevant institutional policies and guidelines, and also recommendations and guidance from expert ethics/advisory committees. It could flag any potential inconsistencies or deviations, prompting the IRB to provide additional justification or to reconsider its approach. It could also flag inconsistencies among the protocol approved by grant funders, the actual protocol submitted for IRB approval, and any descriptions of the protocol contained in informed consent information.
4. *Decision support*: during the deliberation process, IRB members could interact with the LLM—or specifically agentic systems, such as agentic search workflows like Google Gemini's and ChatGPT o3 'Deep Research' features, that have access to both online and institutional resources—to ask for additional information, clarification, or analysis on specific issues. The LLM/agent could quickly retrieve relevant precedents, guidelines, and/or literature to generate a comprehensive, fully cited report to inform the discussion.

DISCUSSION: POTENTIAL BENEFITS, RISKS AND REPLIES

IRB-specific LLMs have the potential to address at least three of the most prominent concerns regarding the use of general LLMs in ethics review.

First, by being trained on and having access to additional information relevant to the specific tasks and context associated with an IRB's work, the accuracy of application-specific LLMs on review tasks is likely to be higher than that of more general models. By drawing on detailed knowledge bases established for the purpose, IRB-specific LLMs (capable of 'reasoning') will be better able to provide relevant information and precedents, and will be less likely to generate inaccurate information.

Second, IRB-specific or expert LLMs may partially ameliorate concerns related to the lack of human empathy and judgement, since such LLMs are trained specifically on past examples of the use of human empathy and judgement in the context of research ethics evaluation.

Third, IRB-specific LLMs are trained on, and have access to, additional information concerning the national or local context of an IRB and its associated institution. This largely addresses concerns based on the need for IRB review to be sensitive to the national or local context in which the IRB operates. IRB-specific LLMs can make use of this information to adjust what would otherwise be context-independent recommendations. Significantly, if important information is missing, this can simply be fed into the vectorised knowledge base as needed.

While IRB-specific LLMs address several key limitations of the more general application of LLMs to IRB review, there are nevertheless multiple remaining issues that need to be acknowledged prior to any actual implementation of these ideas.

Even with careful curation and auditing, there is a risk that the IRB-specific datasets used to train the LLMs could reflect and perpetuate historical inequities or blind spots in the research enterprise. For example, if certain types of research or populations have been systematically underrepresented or excluded from past IRB reviews, the LLM may learn to discount or overlook their ethical significance. Similarly, if the training data contains implicit biases or stereotypes, the LLM may inadvertently reproduce these in its analyses and recommendations.

To mitigate these risks, it will be important to develop robust methods for assessing and mitigating bias in LLM training data and outputs. This could involve techniques such as testing, debiasing, and auditing, as well as ongoing monitoring and adjustment of the models in response to feedback from diverse stakeholders.[36] Post-deployment audits and reinforcement learning from human feedback, based on the differences between

the LLM's reasonings/initial output and final IRB decisions integrating human judgement, are key for continuous improvement.

Another challenge is the potential for over-reliance on LLM outputs in the review process. While the goal is to use LLMs to support and enhance human decision-making, there is a risk that IRB staff and members could become overly deferential to the models' CoT reasonings, analyses, and recommendations, especially if they are presented as or otherwise seen as highly logical, accurate, or authoritative.[37] This could lead to a gradual erosion of the critical thinking and independent judgement that are essential to effective ethical oversight.[38] Moreover, any perception of IRB over-reliance on or deference to LLMs would challenge existing societal expectations that ethical decision-making should be performed exclusively by human persons. While the ultimate aim of deploying such LLMs is to ensure and uphold the ethical conduct of research—which in turn, supports public trust in science—stakeholders (eg, researchers, participants, and the broader public) may ironically perceive their use by IRBs as a disingenuous abrogation of duty, thereby undermining trust and confidence in the research ethics oversight system.

To counter this risk, it will be important to design the LLM-assisted review process in a way that actively engages and empowers human decision-makers. This could involve providing training and support to IRB staff and members on how to write effective prompts to instruct the LLM (ie, user-level prompt engineering) and how to actively and critically interpret and respond to LLMs' CoTs and outputs, and including reminders in the LLM outputs to users that it is not a substitute for human judgement and that the risk of hallucinations is always non-zero. It would also require establishing clear protocols for when and how human judgement should take precedence over machine recommendations. Additionally, in line with emerging norms around disclosing LLM use (eg, in academic writing),[39] IRBs ought to be transparent with stakeholders about their use of such technologies. Relatedly, they may also consider engaging stakeholders—both prior to and following the deployment of LLMs in their review processes—to foster continued trust and confidence in the oversight system.

Notably, even with the accessibility of an LLM's CoT, there still remain important questions about the transparency and explainability of IRB-specific LLMs, and how these factors may impact the perceived legitimacy and trustworthiness of the review process. Given the high stakes and sensitive nature of research ethics, it is desirable that the reasoning behind LLM outputs can be clearly understood and interrogated by human reviewers and other stakeholders.[40 41] This issue can be partially mitigated through (1) application-specific IRB fine-tuning, (2) the extensive use of vectorised knowledge databases (as it would then be clear what information was used to produce an output), and (3) system-level prompt engineering and design settings (eg, low settings of the temperature parameter; instruction to always find factual support based on the RAG knowledge base). Importantly, it could be primarily addressed through the use of systems specifically designed to be explainable, such as reasoners and their CoT. However, this in turn raises the question of how 'raw' or 'full' an LLM's 'thought' processes should be to be considered sufficiently explainable, though this is a topic beyond the scope of this article.

Finally, we note that LLMs may have different utility in reviewing biomedical research protocols than in reviewing social/behavioural protocols. Risk in biomedical studies is often objective and quantifiable (eg, the risk of a device-implantation surgery, or the known side effects of a study drug). Social/behavioural risks are often more subjective—eg, the risk that participation in a given study could harm a participant's reputation, or the level of privacy risk raised by participation in an online focus group. Different IRBs can be expected to decide such questions differently, so the use of LLMs for study screening and risk assessment may not create uniformity across IRBs. On the other hand, the use of an LLM by a given IRB might decrease inconsistency of risk judgements within that IRB over time, as the LLM gains 'experience' through reinforcement learning through human feedback based on the actual decisions. LLMs could also be designed to suggest tools for risk mitigation in social/behavioural studies, such as choosing meeting places where the participant will not be seen meeting the researcher, or providing the researcher's contact information in a format that will not be recognised as 'research-related' by participants' friends and family members, or using a particular approved mechanism for storing sensitive social data in the cloud.

CONCLUSION

Ultimately, the goal of this work is to foster a more ethical, efficient, and responsive system of research oversight that can keep pace with the rapidly evolving landscape of biomedical and social/behavioural research while upholding the highest standards of participant protection and scientific integrity. The background and reasons presented above convince us that IRB-specific LLMs at the very least show significant potential as tools towards the realisation of this goal. Before anyone can know with certainty whether this is the case, much more theoretical and practical work remains to be done, and pilot studies will need to be conducted and thoroughly evaluated. Nevertheless, the reasons laid out above convince us that this work represents a promising area worthy of these additional efforts.

X Jiehao Joel Seah @joelseahjh and Brian D Earp @briandavidearp

Acknowledgements Any use of generative AI in this manuscript adheres to ethical guidelines for use and acknowledgment of generative AI in academic research.[39] Each author has made a substantial contribution to the work, which has been thoroughly vetted for accuracy, and assumes responsibility for the integrity of their contributions.

Contributors SPM, BDE and SJJ conceived the idea. SPM, BDE, SJJ, SL, JS and MA contributed to conceptualisation. SPM, BDE, SJJ, SL, JS and MA contributed to drafting, editing and writing. SPM is the guarantor.

Funding This study was supported by the National Research Foundation Singapore (AISG3-GV-2023-012), Novo Nordisk Fonden (NNF23SA0087056) and Wellcome Trust (226801).

Competing interests SPM sits on the advisory board of AminoChain, a blockchain biobanking startup. JS is a Partner Investigator on an Australian Research Council, grant LP190100841, which involves an industry partnership with Illumina. He does not personally receive any funds from Illumina. JS is a Bioethics Committee consultant for Bayer. The other authors declare no competing interest.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study.

ORCID iDs
Sebastian Porsdam Mann http://orcid.org/0000-0002-1867-2097
Jiehao Joel Seah http://orcid.org/0009-0005-5190-6691
Brian D Earp http://orcid.org/0000-0001-9691-2888

REFERENCES

1 Beecher HK. Ethics and clinical research. *N Engl J Med* 1966;274:1354–60.
2 Gere C. *Pain, pleasure, and the greater good: from the panopticon to the skinner box and beyond*. Chicago: University of Chicago Press, 2017.
3 Arellano AM, Dai W, Wang S, *et al*. Privacy Policy and Technology in Biomedical Data Science. *Annu Rev Biomed Data Sci* 2018;1:115–29.
4 Chisholm A, Askham J. *A review of professional codes and standards for doctors in the UK, USA and Canada*. Oxford: Picker Institute Europe, 2006.
5 International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Integrated addendum to ICH E6(R1): guideline for good clinical practice E6(R2). 2016.
6 Rumbold JMM, Pierscionek BK. A critique of the regulation of data science in healthcare research in the European Union. *BMC Med Ethics* 2017;18:27.
7 World Medical Association. Declaration of Helsinki. 2013.
8 Grady C. Institutional Review Boards: Purpose and Challenges. *Chest* 2015;148:1148–55.
9 Abbott L, Grady C. A systematic review of the empirical literature evaluating IRBs: what we know and what we still need to learn. *J Empir Res Hum Res Ethics* 2011;6:3–19.
10 Lidz CW, Appelbaum PS, Arnold R, *et al*. How closely do institutional review boards follow the common rule? *Acad Med* 2012;87:969–74.
11 Lynch HF. Opening Closed Doors: Promoting IRB Transparency. *J Law Med Ethics* 2018;46:145–58.
12 Sridharan K, Sivaramakrishnan G. Leveraging artificial intelligence to detect ethical concerns in medical research: a case study. *J Med Ethics* 2024.
13 Sridharan K, Sivaramakrishnan G. Assessing the Decision-Making Capabilities of Artificial Intelligence Platforms as Institutional Review Board Members. *J Empir Res Hum Res Ethics* 2024;19:83–91.
14 Muyskens K, Ma Y, Menikoff J, *et al*. When can we Kick (Some) Humans "Out of the Loop"? An Examination of the use of AI in Medical Imaging for Lumbar Spinal Stenosis. *Asian Bioeth Rev* 2025;17:207–23.
15 Haltaufderheide J, Ranisch R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *NPJ Digit Med* 2024;7:183.
16 Porsdam Mann S, Earp BD, Nyholm S, *et al*. Generative AI entails a credit–blame asymmetry. *Nat Mach Intell* 2023;5:472–5.
17 Boroş T, Chivereanu R, Dumitrescu S, *et al*. Fine-tuning and retrieval augmented generation for question answering using affordable large language models. Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP)@ LREC-COLING; 2024:75–82.
18 Whitney SN, Schneider CE. Viewpoint: a method to estimate the cost in lives of ethics board review of biomedical research. *J Intern Med* 2011;269:396–402.
19 Rodriguez E, Pahlevan-Lbrekic C, Larson EL. Facilitating Timely Institutional Review Board Review: Common Issues and Recommendations. *J Empir Res Hum Res Ethics* 2021;16:255–62.
20 Emanuel EJ, Wood A, Fleischman A, *et al*. Oversight of Human Participants Research: Identifying Problems To Evaluate Reform Proposals. *Ann Intern Med* 2004;141:282–91.
21 Seykora A, Coleman C, Rosenfeld SJ, *et al*. Steps toward a System of IRB Precedent: Piloting Approaches to Summarizing IRB Decisions for Future Use. *Ethics & Human Research* 2021;43:2–18.
22 Fernandez Lynch H, Hurley EA, Taylor HA. Responding to the Call to Meaningfully Assess Institutional Review Board Effectiveness. *JAMA* 2023;330:221–2.
23 Bommasani R, Hudson DA, Adeli E, *et al*. On the opportunities and risks of foundation models. *arXiv* [Preprint] 2021.
24 Bubeck S, Chandrasekaran V, Eldan R, *et al*. Sparks of artificial general intelligence: early experiments with gpt-4. *arXiv* [Preprint] 2023.
25 Mumtaz U, Ahmed A, Mumtaz S. LLMs-Healthcare: Current applications and challenges of large language models in various medical specialties. *AIH* 2024;1:16.
26 Meier LJ, Hein A, Diepold K, *et al*. Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept. *Am J Bioeth* 2022;22:4–20.
27 Moradi M, Blagec K, Haberl F, *et al*. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv* [Preprint].
28 Tan Y, Zhang Z, Li M, *et al*. MedChatZH: A tuning LLM for traditional Chinese medicine consultations. *Comput Biol Med* 2024;172:108290.
29 Lewis P, Perez E, Piktus A, *et al*. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv Neural Inf Process Syst* 2020;33:9459–74.
30 Akkiraju R, Xu A, Bora D, *et al*. FACTS about building retrieval augmented generation-based chatbots. *arXiv* [Preprint] 2024.
31 Zohny H, Porsdam Mann S, Earp BD, *et al*. Generative AI and medical ethics: the state of play. *J Med Ethics* 2024;50:75–6.
32 Porsdam Mann S, Earp BD, Møller N, *et al*. AUTOGEN: A Personalized Large Language Model for Academic Enhancement-Ethics and Proof of Principle. *Am J Bioeth* 2023;23:28–41.
33 Porsdam Mann S, Earp BD, Møller N, *et al*. AUTOGEN and the Ethics of Co-Creation with Personalized LLMs-Reply to the Commentaries. *Am J Bioeth* 2024;24:W6–14.
34 Earp BD, Porsdam Mann S, Allen J, *et al*. A Personalized Patient Preference Predictor for Substituted Judgments in Healthcare: Technically Feasible and Ethically Desirable. *Am J Bioeth* 2024;24:13–26.
35 Savulescu J. The structure of ethics review: expert ethics committees and the challenge of voluntary research euthanasia. *J Med Ethics* 2018;44:491–3.
36 Li Y, Du M, Song R, *et al*. Mitigating social biases of pre-trained language models via contrastive self-debiasing with double data augmentation. *Artif Intell* 2024;332:104143.
37 Schwan B. Weighing Patient Preferences: Lessons for a Patient Preferences Predictor. *Am J Bioeth* 2024;24:38–40.
38 Lee H-P, Sarkar A, Tankelevitch L, *et al*. The impact of generative ai on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In Proceedings of the 2025 CHI conference on human factors in computing systems; April 26, 2025:1–22. Available: https://dl.acm.org/doi/proceedings/10.1145/3706598
39 Porsdam Mann S, Vazirani AA, Aboy M, *et al*. Guidelines for ethical use and acknowledgement of large language models in academic writing. *Nat Mach Intell* 2024;6:1272–4.
40 Chu Y, Liu P, Savulescu J, *et al*. n.d. Road Rage Against the Machine: Humans and LLMs Share a Blame Bias Against Driverless Cars. *International Journal of Human–Computer Interaction*:1–11.
41 Porsdam Mann S, Earp BD, Liu P, *et al*. Reasons in the Loop: The Role of Large Language Models in Medical Co-Reasoning. *Am J Bioeth* 2024;24:105–7.